Austrian Journal of Statistics
January 2021, Volume 50, 1–23.
http://www.ajs.or.at/
doi:10.17713/ajs.v50i1.921

A Graphical Tool for Assessing the Suitability of a Count Regression Model

Paul Wilson
University of Wolverhampton

Jochen Einbeck
Durham University

Abstract

Whilst many numeric methods, such as AIC and deviance, exist for assessing or comparing model fit, diagrammatic methods are few. We present here a diagnostic plot, which we refer to as a 'Quantile Band plot', that may be used to visually assess the suitability of a given count data model. In the case of diagnosed model inadequacy, the plot has the unique feature of conveying precise information on the character of the violation, hence pointing the data analyst towards a potentially better model choice.

Keywords: count data regression, goodness-of-fit, Poisson-binomial distribution, mid-quantiles, digit preference.

1. Introduction

Consider univariate count data $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$, possibly accompanied by associated covariate vectors $x_i \in \mathbb{R}^p$, $i = 1. \ldots, n$. Any attempt at drawing inferential conclusions from these observations will require, at first place, a choice of the response distribution, which determines the likelihood function and hence impacts on all further analysis, whether this is done in a Bayesian or frequentist framework. Typical choices for count data distributions would be for instance the Poisson distribution (which describes the number of independent events occurring at constant rate within a certain time interval), the Negative Binomial distribution (which can be considered as an overdispersed extension of the Poisson distribution), the Poisson Inverse Gaussian distribution, the Zero-Inflated Poisson distribution, and many more.

Despite the availability of this plethora of possible count data models, applied users will generally strive for using the most simple model wherever possible, which, in this framework, is the Poisson distribution. The immediate question is then whether a particular choice of distribution is appropriate; which can be phrased as the question of whether the observed data are 'plausible' given the properties of the count data distribution being postulated. One can disassemble this question as asking whether the observed number of zeros is plausible given the postulated count data model (in other words, whether the number of zeros is 'consistent' with the number of zeros predicted by this model), whether the number of ones is plausible given the postulated model, and so on. Clearly, due to the inherent randomness of the system, there will generally be more than one 'number of counts k' which is plausible, and henceforth

the task will be to provide an appropriate range of plausible numbers of counts, as a function of k. The methodology presented in this manuscript will do exactly this, and it will summarize the information, for all count values $k = 0, 1, 2, 3 \ldots$, in a novel diagrammatic tool.

The resulting diagnostic plot may be used to visually assess the suitability of a given count data model. If it is determined that the model in question is not suitable, the plot has the unique feature of conveying precise information on the character of the violation. For instance, this tool will, at a glance, allow the user to detect not only the presence of zero-inflation (more zeros than could be plausibly expected under the given count data model), but also yet poorly explored features such as for instance '3-deflation' (less observations of the count 3 than would be expected under the count data model) or '10-inflation'. This turns out to be highly relevant for assessing the presence of digit preference in a given count data set. The insights gained through this diagnostic plot should also help the data analyst in making a more informed model choice; for instance it is clear that in the case of detected zero-inflation in a Poisson model, a zero-inflated Poisson (ZIP) model should be more adequate.

Denote the postulated response distribution by $G(\theta_i)$, where $\theta_i \in \mathbb{R}^d$ is a multivariate parameter vector, where each of the d components may or may not depend on covariates. It is helpful to think of θ_i as being composed of a mean component, $\mu_i = E(Y_i|x_i)$, with $x_i \in \mathbb{R}^p$, and the remaining (shape, scale, dispersion,...) parameters bundled in δ_i, so that $\theta_i = (\mu_i, \delta_i)$, though it is not strictly necessary that any of the d components actually represents the mean. The δ_i may depend on the same or other covariates as μ_i, but will often be assumed not to depend on covariates at all. In the special case that G is the Poisson or geometric distribution, δ_i is empty.

The question of interest is whether the data $\mathcal{Y}$ are plausible given the distributional assumption G; that is, whether it can be plausibly assumed that $\mathcal{Y}$ have in fact been generated from G. At some occasion, the parameters θ_i may be fixed and known (even if they depend on i), but more often they will be unknown and need to be estimated from the data. We will, initially, not distinguish between these two cases, that is we assume that routines to obtain estimates $\hat{\theta}_i$ are readily available.

We would like to make the point that we do not consider the developed graphical tool as a technique to test for the adequacy of the predictor specification; for instance, of the mean component $\mu_i = E(Y_i|x_i)$. Obviously, this specification is an important part of the model, but the overall distribution of the number of zeros, 1's, etc, will generally not depend strongly on it. Our concern is the suitability of the distribution G, given a certain choice of θ_i.

To fix the notation more precisely, denote $p_i(k) = P(k|\theta_i)$ the probability of observing the count k under covariate x_i and model G, which can be estimated by $\hat{p}_i(k) = P(k|\hat{\theta}_i)$ from the fitted model. For instance, in the special case that $G(\theta_i)$ corresponds to $\text{Pois}(\mu_i)$, one has $\hat{p}_i(k) = \exp(-\hat{\mu}_i)\hat{\mu}_i^k/k!$. This scenario is discussed in Wilson and Einbeck (2018) with focus on the case $k = 0$. This manuscript generalizes those ideas to general k and G and proposes a generic diagrammatic tool.

Denote by $N(k)$ the 'counts of counts', that is the number of occurrences of count k among the data in $\mathcal{Y}$, where $\sum_{k\geq 0} N(k) = n$. It is clear that, for fixed k, $N(k)$ can be described by a sum of n Bernoulli trials with success probabilities $p_i(k)$, $i = 1, \ldots, n$. The resulting distribution, of which one can think as a Binomial distribution with unequal success probabilities, is known as a Poisson–Binomial distribution (Chen and Liu 1997), some properties of which we summarize in Appendix A. Hence, for any choice of k and G, a range of plausible values of $N(k)$ can be obtained from this distribution, using fitted success probabilities $\hat{p}_i(k)$ as model parameters. By doing this for a range of values of k, one can draw diagrams which give envelopes for plausible values of $N(k)$ which can then be compared to the true values. For reasons that will become clear in later sections, we refer to such diagrams as *Quantile Band plots*.

The remainder of this exposition is organized as follows. The graphical tool will be presented and explained in systematic form in Section 2, using an example involving digit preference for illustrative purposes. Computational details of the methodology as well as the problem

of parameter estimation are deferred to Section 3. Further examples, including real data examples, follow in Section 4, before the paper is concluded in Section 5. Some complementary technicalities and definitions are included in the Appendices.

2. Quantile band plots

Based on the principles outlined above, we propose here a diagnostic plot to visually assess the suitability of a given model for the data. We firstly present the algorithm for the construction of the plot.

2.1. Algorithm for plot construction

The first aspect to decide on is the range $K = [k_{min}, k_{max}]$ of count values that is to be assessed. Typical choices would be $k_{min} = 0$ and $k_{max} = \max(\mathcal{Y})$ (and this is what will be used by the default in the graphical tool). Other choices may be preferable in specific circumstances.

A diagnostic plot may be constructed as follows. (The items labelled by a * symbol are to be understood as optional. While the Quantile Band plot as advocated in this work includes the execution of these optional items, there may be certain situation when the data analyst might prefer to omit them; for instance if the quantitative information on the count frequencies is to be conveyed through the plot.)

Specification Determine the model $G(\theta_i)$ for the data $\mathcal{Y}$. Obtain estimates $\hat{\theta}_i$, $i = 1, \ldots, n$ where required.

Computation For k in K

(i) compute $\hat{p}_i(k) = P(k|\hat{\theta}_i)$;

(ii) from the Poisson-Binomial distribution with parameters $\hat{p}_i(k)$, $i = 1, \ldots, n$, compute lower and upper quantiles $q_{\alpha/2}(k)$ and $q_{1-\alpha/2}(k)$;

*(iii) compute also the median, $m(k)$, and use it to compute shifted versions $A(k) = N(k) - m(k)$, $b_\gamma(k) = q_\gamma(k) - m(k)$, for $\gamma \in \{\alpha/2, 1 - \alpha/2\}$.

Create graph

(i) Plot the functions $b_{\alpha/2}(k)$ and $b_{1-\alpha/2}(k)$ versus k. Then add to the plot the observed shifted counts, $A(k)$, of the observed data $\mathcal{Y}$. [If item (iii) above has not been carried out, replace b and A by q and N, respectively, and in this case one may optionally add $m(k)$ to the plot.]

*(ii) Rotate the plot by 90 degrees, so that k is orientated along the vertical axis.

If the data is consistent with the distribution fitted, the curve $A(k)$ [$N(k)$, respectively] should (largely) stay within the bands $b_{\alpha/2}(k)$ and $b_{1-\alpha/2}(k)$ [$q_{\alpha/2}(k)$ and $q_{1-\alpha/2}(k)$, respectively]. If the data is *not* consistent with the distribution fitted then $A(k)$ [$N(k)$] is likely not to stay within these bands. Informally we will refer to the line representing the $b(k)$'s as the upper and lower quantile bands, and the line representing the $A(k)$'s as the count-line.

Note that there are several possible choices of how exactly to compute quantiles for a discrete distribution. For our purposes, the quantiles employed are the *mid-quantiles*, which are discussed in Section 3.1.

2.2. Illustration via digit preference

Digit Preference refers to the mis-reporting of some numbers in favour of "preferred numbers". An early example in the literature is Myers (2002) who reports a tendency in the the 1910, 1920, and 1930 U.S. censuses to report ages of 20 as 21 and ages of 31 as 30. Camarda, Eilers, and Gampe (2002) provide an extensive list of references to literature concerning digit preference, and report that maybe the most common form of the phenomenon is "heaping" of data at multiples of 5.

We present here a hypothetical example that simulates a situation where the number of, say, annual theatre visits follows a Poisson distribution with parameter 7, however in practice the following occurs:

- visit counts of 0, 1 and 2 are correctly reported;
- visit counts that are a multiple of 5 are correctly reported;
- visit counts that are within 2 of a multiple of 5 are reported correctly with probability ϕ and as that multiple of 5 with probability $1 - \phi$.

Table 1: Reported numbers of theatre visits

Annual Visits	0	1	2	3	4	5	6	7	8
Frequency	1	5	9	20	32	106	69	64	57
Annual Visits	9	10	11	12	13	14	15	16	
Frequency	35	55	22	9	4	4	6	2	

Table 1 presents a sample of 500 such data, where 20% of non-multiples of 5 are mis-reported as the nearest multiple of 5 (i.e. $\phi = 0.80$). Note that the mean and variance of these data are 7.01 and 7.33 respectively, thus, based solely on these statistics, a Poisson model appears reasonable.

Table 2: Data of Table 1 with upper and lower quantiles for $N(k)$ and $A(k)$ (first 8 rows).

k	$N(k)$	$q_{0.025}(k)$	$q_{0.975}(k)$	$m(k)$	$A(k)$	$b_{0.025}(k)$	$b_{0.975}(k)$
0	1	0.00	2.49	0.39	0.61	-0.39	2.10
1	5	0.05	7.21	3.01	1.99	-2.96	4.20
2	9	5.05	18.05	10.95	-1.95	-5.90	7.10
3	20	16.58	36.13	25.80	-5.80	-9.22	10.33
4	32	33.20	58.49	45.32	-13.32	-12.12	13.17
5	106	49.40	78.72	63.59	42.41	-14.19	15.13
6	69	59.12	90.39	74.30	-5.30	-15.18	16.09

The question of interest is whether the data could plausibly have been generated from a Poisson distribution. Therefore, the above procedure is applied to the data of Table 1, choosing $K = [0, 16]$ as the range of counts of interest, and omitting the two 'starred' items for now.

The values of $N(k)$, $q_{0.025}(k)$ and $q_{0.975}(k)$, for $k = 0, 1, \ldots, 7$ are given in Table 2; the full table is given in Table 9 of Appendix D. From the left–hand part of this table one can construct the diagnostic plot displayed in Figure 1. Clearly $N(5) > q_{0.975}(5)$ and $N(10) > q_{0.975}(10)$ as is illustrated by the red line which represents the $N(k)$ lying above the green line representing the $q_{0.975}(k)$ for $k = 5$ and $k = 10$. Similarly the red line representing the $N(k)$ lies beneath the green line representing the $q_{0.025}(k)$ for some other values of k, indicating the unsuitability of a Poisson model here.

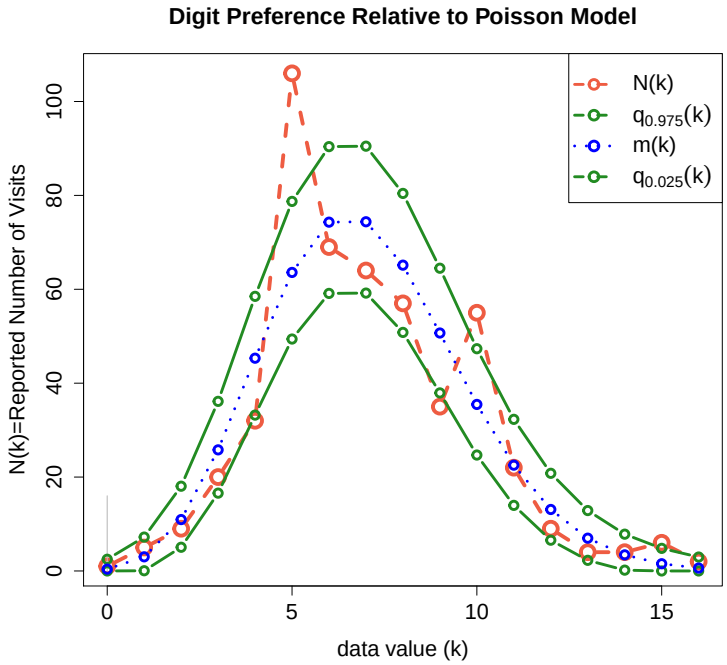

Figure 1: Diagnostic plot for the data of Table 1. We refer to plots in this form as *raw* Quantile Band plots.

Whilst the plot of Figure 1 contains complete information, the range of $N(k)$ leads to the bands described by $N(k)$, $q_{0.025}(k)$ and $q_{0.975}(k)$ being somewhat close. For data where the range of $N(k)$ is large, it frequently occurs that these bands become extremely close and difficult to distinguish (examples of such plots are given in Figure 9 in Appendix E).

A superior plot may be obtained by including the 'starred items' from Section 2.1. After subtracting the median from all other quantities we arrive at the information displayed in the right part of Table 2, and hence the Quantile Band plot in the top panel of Figure 2. Similar to the display of boxplots, which can be presented either vertically or horizontally, one may display the Quantile Band plot in either orientation. We prefer the vertically rotated version, as shown in the bottom panel of Figure 2, and will use this version throughout the remainder of this paper. We acknowledge however that others may prefer the horizontal version, and also that there are occasions when practical considerations such as availability of space in a publication render the horizontal version preferable. One immediately concludes from this plot that there is evidence of inflation of multiples of five, and also some evidence of deflation of the counts four and nine (which is arguably an artifact of the former). Overall, this gives evidence that the Poisson assumption is not adequate. We further discuss this example in Section 4.1.

Enhanced interpretations on other diagnostic aspects can often be drawn from the specific nature of the pattern; these will be discussed in the examples of Section 4. We will turn now to the question of how exactly the quantiles of the distribution of $N(k)$ are obtained.

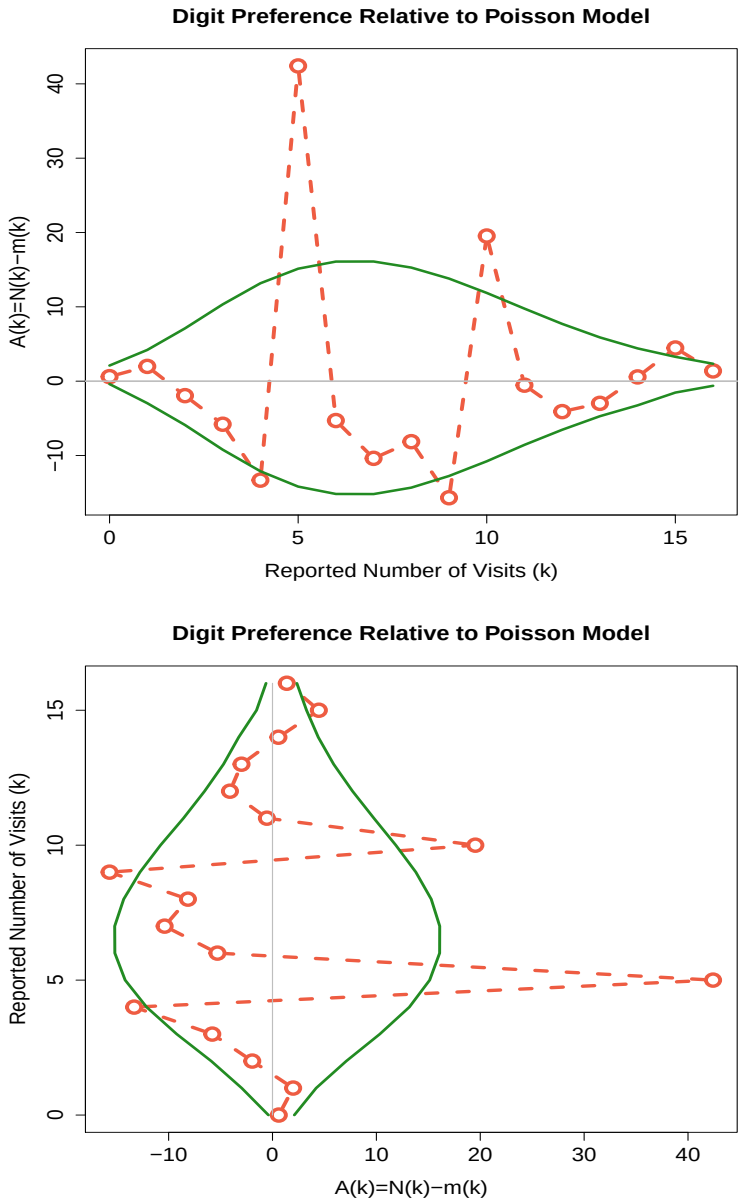

Figure 2: Construction of Quantile Band plots for simulated covariate-free data. Subtracting the blue coloured median curve $m(k)$ in the plot of Figure 1 from all other curves gives the horizontal version of the plot (top), which after rotation leads to the vertical Quantile Band plot (bottom).

3. Technical details

3.1. Quantiles and mid-quantiles

We consider now the question of how exactly to compute the quantiles of the distribution of $N(k)$. Recall that for each (fixed) value of k, $N(k)$ is described as a Poisson-Binomial distribution with parameters $\hat{p}_i(k)$, $i = 1, \ldots, n$. We denote by F the distribution function of $N(k)$, that is $F(t) = P(N(k) \leq t)$. The γ–quantile of this distribution is traditionally defined as

$$Q_{\text{trad}}(\gamma) = \min_t\{F(t) \geq \gamma\},$$

where t is a non–negative real value. For the purposes of Section 2 one will have either $\gamma = \alpha/2$ or $\gamma = 1 - \alpha/2$. For the Poisson-Binomial distribution, these quantiles can be obtained straightforwardly from the package poibin (Hong 2013). However, as Ma, Genton, and Parzen (2011) have pointed out, there are "many drawbacks" with using traditional quantiles for discrete distributions. Firstly, the quantile function is discontinuous, leading to unfavourable theoretical properties. Secondly, it causes interpretational problems: For instance, for all observable values $t' \in \mathcal{N} = \{0, \ldots n\}$ of $N(k)$, the sum of the (right-handed) p-value and the left–handed quantile turns out to be larger than 1, since the probability mass at t' is 'double-counted'.

The suggested improvement by Ma *et al.* (2011) is to split the probability mass at discrete values accordingly, which gives rise to the definition of the mid-distribution function $F_{\text{mid}}(t) = F(t) - 0.5p(t)$, where $p(t)$ is the corresponding point mass at t (which is equal to 0 for all $t \notin \mathcal{N}$).

This concept then leads to the definition of mid-quantiles, and is also naturally related to the notion of mid-p-values (Franck 1986). The mid-quantile–function, which we denote by $Q_{\text{mid}}(\gamma)$ henceforth, is constructed so that, at the points of observable values t', one has $Q_{\text{mid}}(F_{\text{mid}}(t')) = t'$, and a piecewise linear interpolation in between (Ma et al, 2011). Since the exact mathematical expression of mid-quantiles in itself is rather complicated, we have deferred this alongside with a tutorial to Appendix B.

In all computations leading to Quantile Band plots in this paper, we have used for step (ii) in Section 2 the convention $q_\gamma(k) = Q_{\text{mid}}(\gamma)$, with the distribution F which gives rise to this quantile being the distribution of $N(k)$ corresponding to an underlying count distribution G. Of course, also the medians $m(k)$ are computed as 'mid–medians' in this manner. Wilson and Einbeck (2018) constructed mid-quantiles in the special case $k = 0$ and $G = \text{Pois}$, and referred to the resulting intervals $[Q_{\text{mid}}(\alpha/2), Q_{\text{mid}}(1-\alpha/2))$ as *mid-quantile intervals* (MQI). In this spirit, we will also refer to tables which provide information as in the left half of Table 2 as *mid-quantile tables*.

3.2. Parameter estimation

The view taken in our approach is that the production of the Quantile Band plot succeeds the parameter estimation. In other words, it is assumed that, once a user has specified a certain count data model $G(\theta_i)$, a routine to estimate θ_i is readily available. The distribution G and the estimated $\hat{\theta}_i$ then serve as input to the production of the Quantile Band plot. In this sense, the estimation of the θ_i is not considered as an intrinsic part of the methodology as such. Some words on parameter estimation still appear in order.

Usually, the $\hat{\theta}_i$ will be estimated through Maximum Likelihood (ML). A frequent scenario is where $\theta_i = (\mu_i, \delta_i)$, where δ_i is some dispersion or shape parameter. This would be the case, for instance, for the Negative Binomial (Type I or II), Neyman Type A, the Pòlya-Aeppli, or the Poisson-Inverse Gauss distribution. For, say, the latter case one may consider a model with constant dispersion index $\delta_i \equiv \delta$, and a log-link for the mean regression parameter, i.e. $\log(\mu_i) = x_i^T\beta$, or equivalently $\mu_i = E(Y_i|x_i) = \exp(x_i^T\beta)$, for some vector of predictors $x_i \in$

$\mathbb{R}^p$. The actual vector of parameters to estimate would then be $(\beta, \delta) \in \mathbb{R}^{p+1}$. But of course, it is also possible to describe the dispersion or other parameters through appropriate linear predictor terms, which may involve the same or different covariates as the mean function. Having obtained the underlying estimates $(\hat{\beta}, \hat{\delta})$, one can immediately obtain the estimates $\hat{\theta}_i$ via the pre-specified predictor configurations, which then allows production of the $\hat{p}_i(k) = P(k|\hat{\theta}_i)$, and hence execution of the machinery outlined in Section 2.

The actual calculation of the involved MLE's will usually be carried out through software. For instance, in the case that G corresponds to a Poisson or Binomial distribution, in which cases $d = 1$, estimation of θ_i reduces to fitting a generalized linear model (McCullagh and Nelder 1989) and hence can be carried out using the `glm` function in R. A wide range of further multi–parameter count data distributions can be fitted using functionalities provided by the R packages **VGAM** (Yee 2010) and **gamlss** (Rigby and Stasinopoulos 2005). Estimation routines for some further count distributions, including the ones mentioned earlier in this subsection, which are relevant for specific applications for instance in dosimetry, are provided in form of R code in the supplementary material of Oliveira, Einbeck, Higueras, Ainsbury, Puig, and Rothkamm (2016).

However, it needs to be pointed out at this occasion that Maximum Likelihood is not the only way to estimate model parameters. For instance, it is long known that the ML estimate of the Poisson mean, that is, the whole sample mean, can perform very poorly if the data are zero-inflated or zero-deflated (or, to condense these two terms, 'zero-modified', see also da Silva, Ribeiro, Conceiçáo, Andrade, and Louzada (2018)). In response to such problems, Plackett (1953), Irwin (1959) and Ridout and C.B. (1992) present formulae for estimating the Poisson parameter from the mean of the positive data values; that is from the zero-truncated data. These estimators reduce the bias of the ML estimator of the mean parameter, but turn out to be less precise than the maximum likelihood estimator. Wilson and Einbeck (2018) went one step further and suggested (for scenarios in which zero-modification is expected or suspected) to balance bias and variance of the Poisson mean estimate through a weighted mean of the whole sample estimator and the zero-truncated estimator, with a weight of 2/3 for the whole sample mean performing favorably in simulation studies. In the case of actual zero-modification, this 'hybrid' estimator will considerably reduce the bias, but in its absence it will not behave notably differently than the whole sample estimator. It is emphasized that the problem being solved through such measures is intrinsic to *zero* counts. For inflation or deflation of higher counts k, the impact on the estimates of μ_i is much less severe since the effects tend to cancel out over neighbouring counts. For multi-parameter distributions, the second parameter can absorb the overdispersion created through the excess zeros to some extent. Hence, in line with this reasoning, we use the hybrid estimation technique in the applications in Section 4 only for one–parameter distributions, i.e. the Poisson and the geometric distribution. The example in Section 4.3 will illustrate the impact of the different mean estimators on the Quantile Band plot explicitly.

3.3. Multiple testing issues

One may argue that due to the consideration of a sequence of mid-quantile intervals for $N(0), N(1), \ldots$ one has to account for multiple testing issues. It is certainly true that the count-line being outside of the mid-quantile bands at $N(0)$ is equivalent to determining that there is zero-modification relative to the null model, as in the test of Wilson and Einbeck (2018), similar statements being possible for other $N(k)$'s, and that if several such tests were performed simultaneously then multiple testing issues would arise. The pragmatic view is that our proposed plot should not be considered as a testing procedure, but as a simple diagrammatic tool which supports the data analyst in identifying potential model inadequacies, similar in spirit to a QQ plot.

4. Applications

4.1. Digit preference revisited

We saw in Section 2.2 that a Poisson model is unsuitable for the data of Table 1. Whilst in the introduction to that example the data generating mechanism is informally described, even if this had not been the case the Quantile Band plots of Figure 2 clearly illustrate that the observed numbers of multiples of five are considerably greater than is to be expected under a Poisson model, and most other values somewhat less than would be expected, which would lead the analyst to suspect that digit preference for multiples of five is a feature of the data.

When the role of the distribution G is taken by a Poisson model that incorporates digit preference ($\phi = 0.8$), that is a modified Poisson model with probability density function g as given in Appendix C, one obtains the Quantile Band plot displayed in Figure 3. We note that for $k > 3$ the quantile bands are wider at points corresponding to multiples of five, and narrower at other points, reflecting the modified densities of the Poisson model which incorporates digit preference. The count-line is now interior to the quantile bands at all points, indicating that $b_{0.025}(k) < A(k) < b_{0.975}(k)$, for all values of k, and hence the suitability of the modified Poisson model.

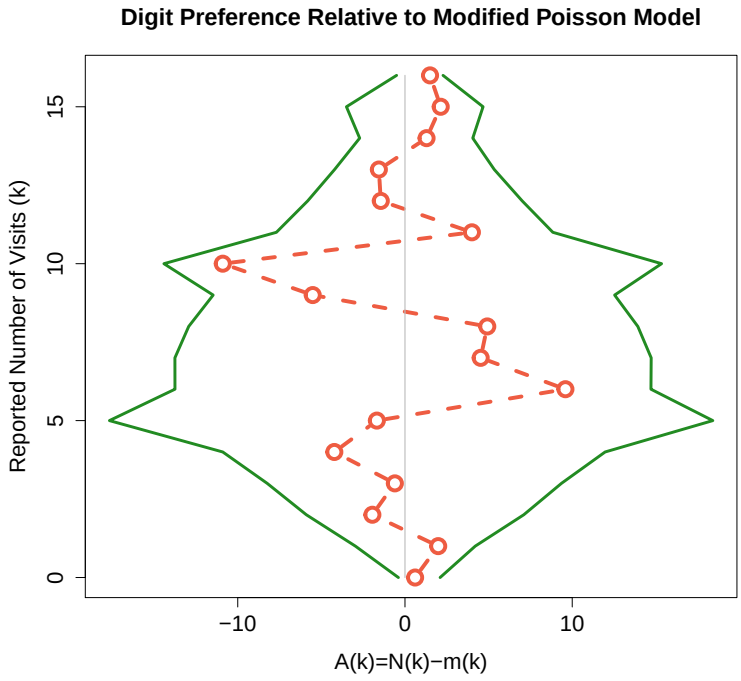

Figure 3: Quantile Band plot for modified Poisson model (incorporating digit preference)

The mid-quantile table corresponding to the Poisson model modified for digit preference is given in Table 10, Appendix D.

4.2. Frequency of homicides

Agresti (2002) discusses the fitting of a variety of models to data concerning the responses to the question: "Within the last twelve months, how many people have you known personally that were victims of homicide?" The respondents were classified as *Black* or *White* according to their race. The data is summarised here in Table 3.

The relative fits of Geometric, Poisson, Poisson Inverse Gaussian (PIG), Negative Binomial (Type I), Negative Binomial (Type II) and Zero-Inflated Poisson (ZIP) models are shown

Table 3: Number of victims of murder known in past year, by race

Response	0	1	2	3	4	5	6
Black	119	16	12	7	3	2	0
White	1070	60	14	4	0	0	1

in Table 4. (For the Geometric and Poisson models the mean is modelled by *race* using a log-link, for the PIG and negative binomial models both parameters are modelled by *race* using log-links, and for for the ZIP model, the mean and zero-inflation parameters are both modelled by *race* using log and logit-links respectively.)

Table 4: Values of information criteria for several count regression models fitted to homicide data.

Model	AIC	BIC
Poisson	1942.87	1953.23
Geometric	3115.99	3126.34
PIG	1007.20	1027.91
NBI	1000.56	1021.26
NBII	1000.56	1021.26
ZIP	998.74	1019.44

Quantile Band plots corresponding to the various models are illustrated in Figure 4, with the corresponding mid-quantile tables provided in Appendix D (Tables 11 to 16). It is apparent that the one parameter geometric and Poisson models are inadequate; we do see however from the plots for the Poisson and geometric models that whilst both underestimate the numbers of zeros, and overestimate the number of 1's, the over-estimation of 1's is not as severe for the geometric model as it is with the Poisson, but the over estimation of larger values is more severe under the geometric model than under the Poisson. It is most interesting to compare the two-parameter models. In all cases the observed counts all fall within their respective mid-quantile intervals, so it may be argued that all four are suitable. Reflecting its position as the poorest of the two parameter models considered here by both AIC and BIC criteria, it is noticeable that the PIG model considerably overestimates the number of 1's, and considerably underestimates the number of 2's and 3's. The Quantile Band plots for the two types of negative binomial are identical; it is notable that the fits of the negative binomial models and the zero-inflated Poisson models are similar under both the AIC and BIC criteria, it is apparent however that the ZIP model slightly underestimates the numbers of 1's, and overestimates the numbers of 2's compared to the observed data, the reverse being the case for the negative binomial models.

We include in Figure 9, Appendix E the mid-quantile band plots obtained when Poisson and PIG models are fitted to the data of Table 3. These plots illustrate the superiority of the median–adjusted plots of Figure 4.

4.3. Choosing between Poisson and ZIP models

In Section 3.2 we discussed the estimation of model parameters, and mentioned three estimators of the parameter of a Poisson model that have been proposed in the literature. We present here an example that further explores this issue, and illustrates the usefulness of Quantile Band plots as a tool in determining the appropriateness of a given model, under consideration of different possible estimates of the model parameters.

Consider the data of Table 5, which are generated by concatenating 200 zeros to a sample of

800 data drawn from a Pois(0.5) distribution, thus the data are zero-inflated by construction. The mean of these data is 0.435 and the variance 0.484, indicating at first glance that a Poisson model may not be unreasonable.

Table 5: Simulated data with zero-inflation

0	1	2	3	4	5
663	256	67	12	1	1

This is now a situation as touched upon in Section 3.2, where the Poisson mean estimate is possibly affected by the presence of zero-inflation. Hence, we consider in this example the application of three different estimators of the Poisson mean, which are the whole sample mean ($\hat{\mu}_W$), zero-truncated mean ($\hat{\mu}_T$), and hybrid mean ($\hat{\mu}_H$), respectively. Table 6 gives the corresponding three estimates along with the log–likelihoods of the Poisson models when using those estimates. It is apparent that, under the log-likelihood criterion, the Pois($\hat{\mu}_W$) model has the best fit (indeed, as $\hat{\mu}_W$ is the maximum likelihood estimator this must be the case), the fit of the Pois($\hat{\mu}_H$) model is only slightly poorer, and that of the Pois($\hat{\mu}_T$) model more so. The log-likelihood statistics on their own however do not say anything about the suitability of the models: it is possible that all or none are compatible with the observed data.

Table 6: Model fits using whole sample mean ($\hat{\mu}_W$), zero-truncated mean ($\hat{\mu}_T$), and hybrid mean ($\hat{\mu}_H$), respectively.

$\hat{\mu}$	log–likelihood
$\hat{\mu}_W = 0.435$	-873.0
$\hat{\mu}_T = 0.534$	-882.9
$\hat{\mu}_H = 0.468$	-874.2

Therefore, we proceed with the production of Quantile Band plots, which are depicted for the three Poisson models as well as the ZIP model in Figure 5. The two plots in the top row show, interestingly, that both a Poisson model with mean parameter of 0.435 and a ZIP model are compatible with the data. What has happened here is that the presence of extra zeros has reduced the value of $\hat{\mu}_W$ (the whole sample mean) to a value which renders the numbers of observed zeros and 1's compatible with a Pois($\hat{\mu}_W$) model. There is no contradiction here, and both statements are correct: a zero-inflated model with small mean parameter is simply hard to distinguish from a Poisson model with even smaller mean parameter (note that the Poisson mean estimate under the ZIP model is 0.533 and the estimated zero-inflation parameter is 0.184).

The bottom plots show that Poisson models with larger mean parameters, as obtained through the truncated and hybrid estimators, are *not* compatible with the data. Also these are true statements: The plots simply assert that Poisson models with means of 0.468 and 0.534, respectively, are incompatible with the data. However, arising from this is the interesting question which of the three Poisson-based plots tells the most meaningful story in terms of the data generating mechanism. From this angle, the hybrid and the truncated estimators do the more useful job, by producing Poisson mean estimates which are closer to the true value (of 0.5), allowing for the indication of the presence of inflation of zeros relative to the Poisson model.

Finally, it is worth noting that, while the count-line of the top-left diagram of Figure 5 corresponding to the Pois($\hat{\mu}_W$) model remains within the quantile bands $b_{0.025}$ and $b_{0.975}$, it does hint that somewhat more zeros and considerably less 1's than expected under the Poisson model are present in the observed data. This pattern drawn by the count-line is a characteristic of zero-inflation, which may lead the analyst to speculate that the data is in fact

Table 7: Frequency of dicentric chromosomes after acute homogeneous *in vitro* exposure to doses between 0 and 4.5Gy of Cobalt-60 γ-rays. (This corresponds to data set A1 in the notation of Oliveira *et al.* (2016), where also the reference for the data source is provided.)

	Frequency of counts					
dose	0	1	2	3	4	5
0.00	2591	1	0	0	0	0
0.25	2185	8	0	0	0	0
0.75	2550	44	1	0	0	0
1.00	2231	54	2	0	0	0
1.50	1712	96	3	0	0	0
2.50	1196	123	7	1	0	0
3.00	1070	320	41	6	1	0
4.50	895	360	110	25	5	1

zero-inflated, and hence to proceed with fitting a ZIP model which then delivers the nicely behaved Quantile Band plot as in the top right panel.

4.4. Biodosimetry data

We consider data consisting of $n = 14430$ chromosome aberration counts previously studied by Oliveira *et al.* (2016). The covariate *dose*, with values between 0 and 4.5Gy, gives the radiation dose applied *in vitro* to blood sample cells, causing DNA damage in form of double-strand breaks. When incorrectly repaired by the cellular DNA-damage response mechanism, this can lead to dicentric chromosomes which can be counted under a microscope. That is, each examined blood sample cell contributes, for known covariate dose, exactly one count observation. For this data set, the counts take values in the range from 0 to 5. Data of this type have been fitted traditionally through Poisson regression models, though deviation from the Poisson property, and specifically the presence of excess zero counts, has been regularly reported in the literature, see e.g. Puig and Barquineiro (2011).

Table 7 displays the data under investigation, and Figure 6 contains the Quantile Band plots obtained when Poisson and zero-inflated Poisson models, using a log–link and quadratic polynomial for dose, but constant zero-inflation parameter in the ZIP case, are fitted to these data. The left hand plot clearly indicates the unsuitability of the Poisson model, whereas the right hand plot indicates that ZIP is suitable.

Oliveira *et al.* (2016) carried out an extensive analysis of this data set, applying several statistical tests and model selection criteria in order to decide for an adequate modelling strategy. Specifically, they found that a Negative Binomial type 2 (hereafter NB2) model returned the lowest AIC (7489.1), closely followed by a ZIP model (AIC=7490.4). Other models considered included the Poisson as reference model (AIC=7504.7), and a Poisson Inverse Gaussian (PIG; AIC=7495.2).

The two plots in Figure 7 corresponding to the NB2 and PIG models, respectively, illustrate cases where the adjusted observed data line, $A(k)$, remains close to the centre line. For the NB2, all observations lie between the 43% and 57% quantiles of their respective Poisson–Binomial distribution. Hence, there is less random variation amongst observed counts than would be expected under the NB2 model, most likely indicating that the variance of the fitted model is inflated in order to accommodate the number of observed zeros. A similar effect is observed for the PIG model. In summary, these plots suggest that the ZIP model is the most adequate model for these data, deviating from what would be concluded by looking at a single–number model selection criterion such as AIC.

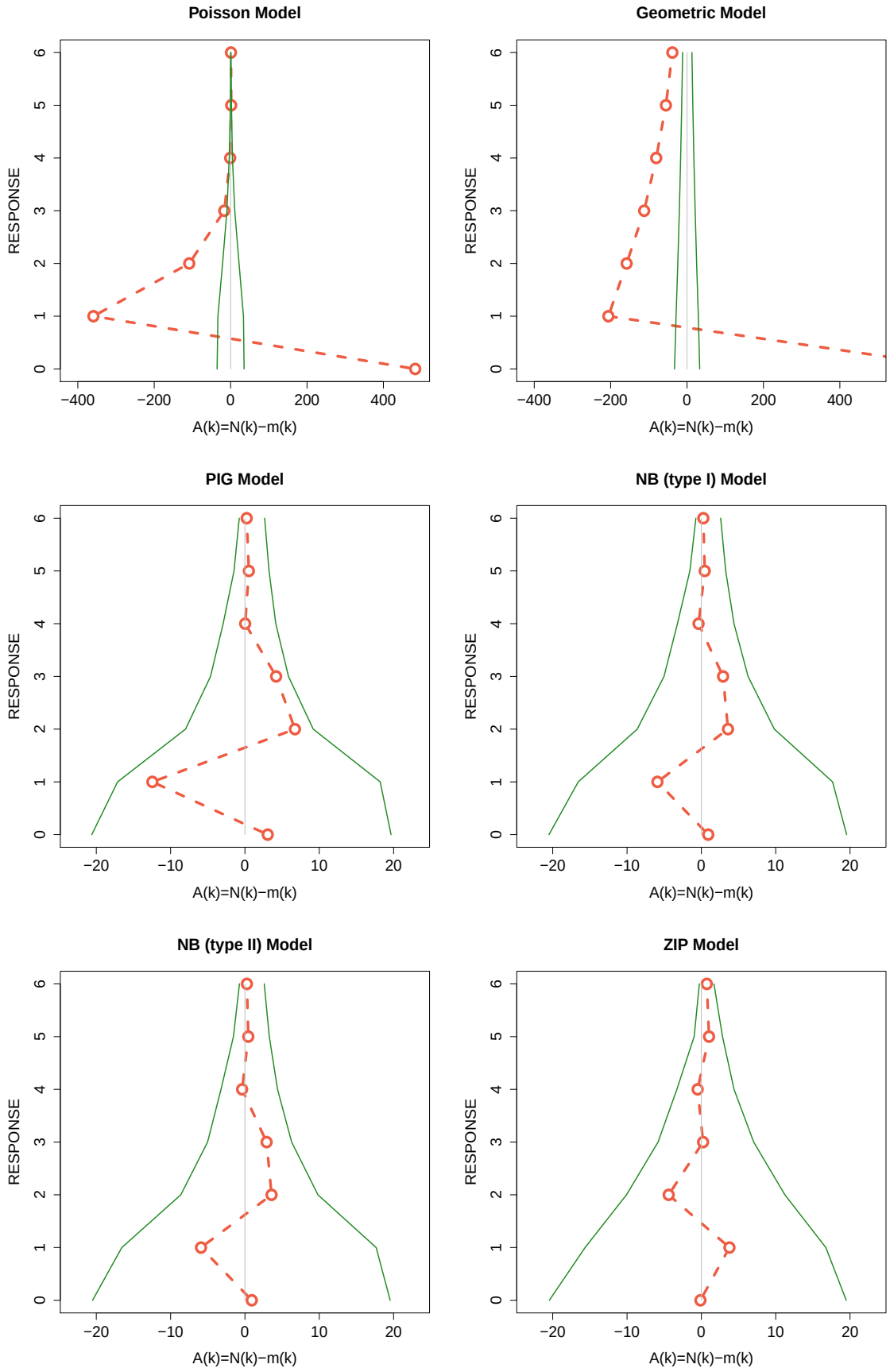

Figure 4: Quantile Band Plots for the homicide data (Table 3). For ease of comparisons, the horizontal axes for the one parameter Poisson and geometric models are drawn to the same scale, and the horizontal axes for the four two-parameter models are drawn to the same scale.

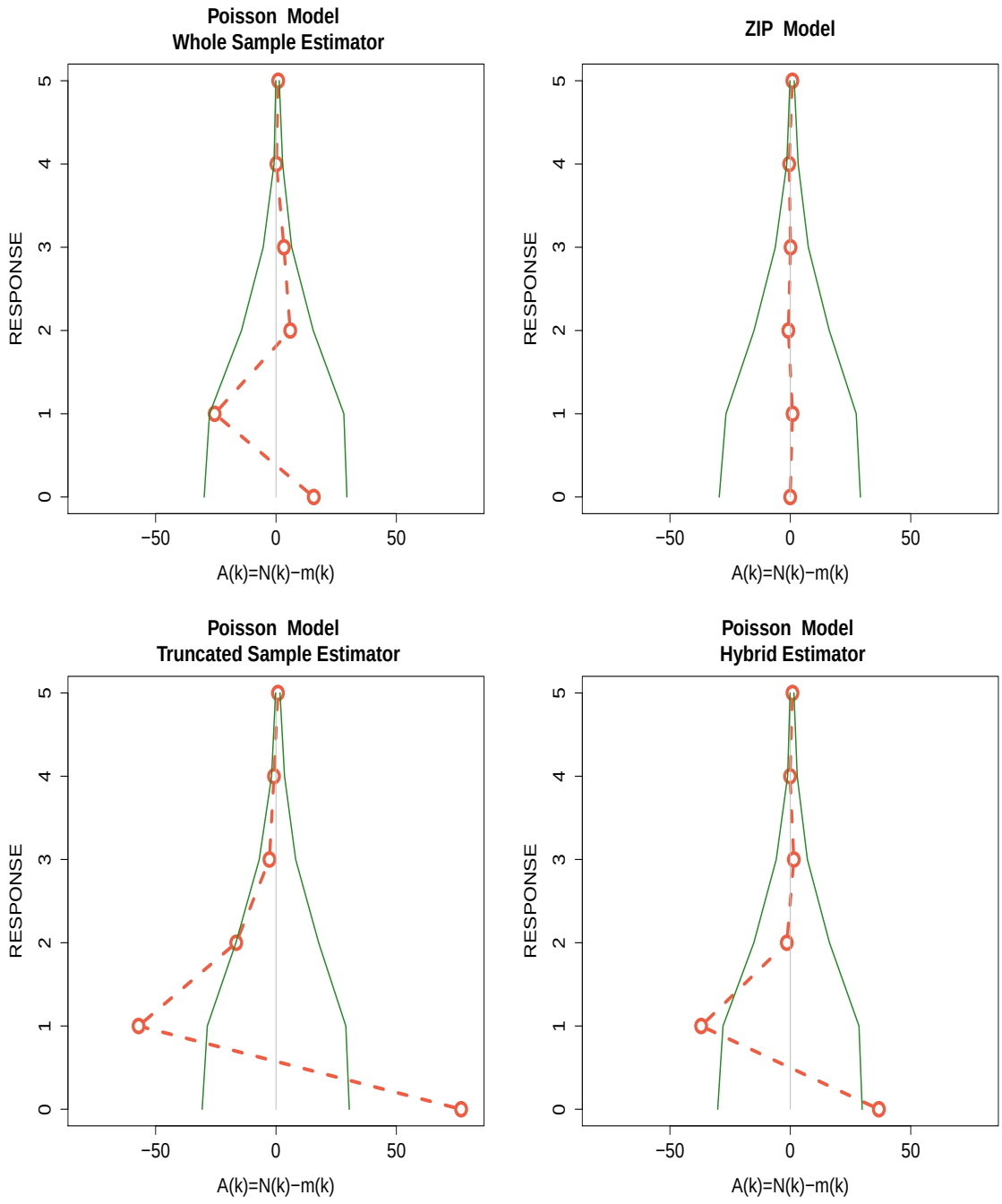

Figure 5: Quantile Band plots for the simulated data of Table 5.

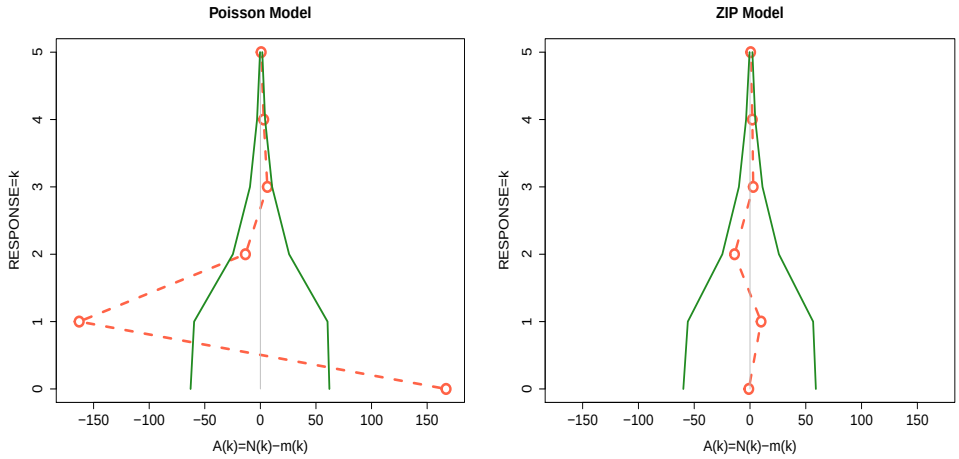

Figure 6: Quantile Band plots for biodosimetry data, with the hypothesized distribution G corresponding to Poisson and ZIP, respectively.

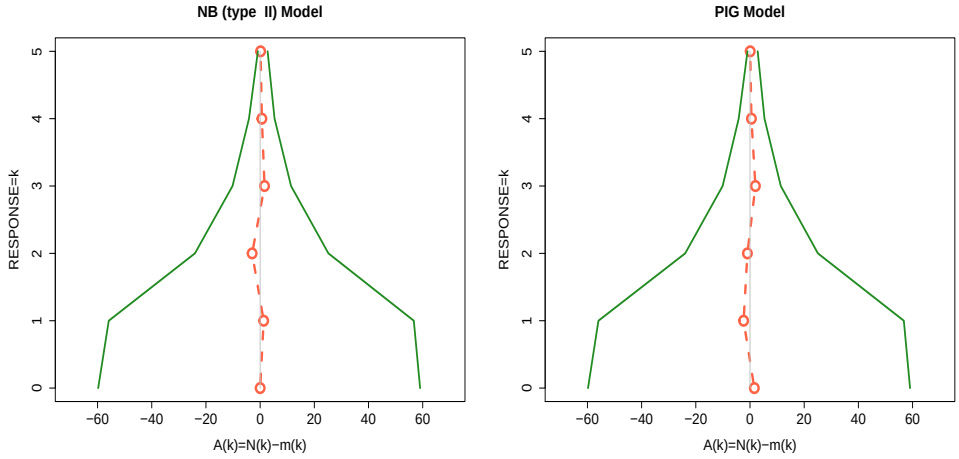

Figure 7: Quantile Band plots for biodosimetry data, with the hypothesized distribution G corresponding to NB2 and PIG, respectively.

5. Conclusion

Whilst the principal purpose of the Quantile Band plots presented here is, as indicated in the title of this paper, to assess the suitability of a count regression model to a given data set, not to determine the 'best' model, the plots are an extremely useful tool to help answer the question: 'what is the best model?'. It is sometimes overlooked that the purpose of fitting a model to a given sample of data is *not* to find the model that 'best fits' the *sample*, but to attempt to discover a model that is, to paraphrase George Box, probably incorrect but of use as a model for the data from which the sample was taken. Whilst likelihood–based methods such as log–likelihood, AIC, BIC and various other 'information criteria' have an enormous role to play in determining and estimating the coefficients of 'the most suitable model', the fact that they frequently disagree on the nature of that model illustrates that none of them are perfect arbitrators. The graphical tool presented in this paper is an alternative approach to assessing model fit, which may be used on its own, or in conjunction with other methods. Its unique feature is that it enables the user to determine whether the observed frequency of a given count in the data is compatible with that to be expected under a given distribution; as stated in Section 2 if the majority of the frequencies of observed counts are indicated as being compatible with the model under consideration (i.e. they lie within the mid-quantile bands), then the model is likely to be appropriate for the data. Of course, under this procedure several models may be deemed 'appropriate'. This is well illustrated by Figure 4 which indicates that all four of the two parameter models considered are appropriate. A strength of the plots is that if they indicate non-suitability of a model, they also indicate the nature of the unsuitability, for example the top two plots of Figure 4 show that Poisson or geometric models are unsuitable as they severely underestimate the amounts of zeros, and overestimate the amount of 1's. Exactly the same information is contained in the horizontal and vertical forms of the plot. Raw Quantile Band plots (for example Figure 1) actually contain more information than the median-adjusted forms, but frequently are impractical (see Figure 9).

Quantile Band plots contain more information however than whether the observed counts are compatible with the stated model. As discussed in the various examples of Section 4 they indicate which values of k are over- or underestimated by the model; the example of Section 4.3 shows how the plots may be used to help determine the most suitable parameter for a model, which may differ from the maximum likelihood estimate. In the example of Section 4.4 the count-lines of the Quantile Band plots for the negative binomial and PIG models of Figure 7 indicate very little variation of the observed counts about the median values, possibly indicating that the counts exhibit less random variation than expected under these models, leading the researcher to speculate as to the reason for the possibly larger than necessary estimates of the dispersion parameter.

Whilst the authors advise the use of mid-quantiles as outlined in Section 3.1 for the construction of the quantile bands, other forms of quantiles may also be adopted. An attractive alternative may also be the use of expectiles, especially with view to their uniqueness property for discrete data (Eilers 2013). At present the plots are only proposed for use with univariate count models, but could be extended to continuous models by binning data, or to multivariate models by increasing the dimension of the plots.

R Code for the production of Quantile Band plots (in all three versions; that is with or without the starred items from Section 2.1) will be made available as supplementary material accompanying this publication.

Acknowledgements

Parts of this paper were written while the first author visited Durham University through the Visitor Programme of the Department of Mathematical Sciences; the first author also received support from the statistical cybermetrics research group and the School of Mathematics and Computer Science of the University of Wolverhampton for this visit. The second author was

partly supported by CRoNoS COST Action IC1408. The authors are grateful to a referee and the Editor for insightful comments.

Appendices

The Poisson-binomial distribution

Let $X_1, X_2, \ldots, X_n$ be n independent Bernoulli random variables with parameters $p_1, p_2, \ldots, p_n$, respectively, and $S = \sum_{i=1}^{n} X_i$. The distribution of S is known as a *Poisson-Binomial* distribution (Chen and Liu, 1997), with probability mass function

$$P(S = s) = \left\{ \prod_{i=1}^{n} (1 - p_i) \right\} \sum_{i_1 < \cdots < i_s} w_{i_1} \cdots w_{i_s} \tag{1}$$

where $w_i = \frac{p_i}{1-p_i}$, $i = 1, 2, \ldots, n$, and the summation is over all possible combinations of distinct $i_1, i_2, \ldots, i_s$ from $\{1, 2, \ldots, n\}$. It has the properties $E(S) = \sum_{i=1}^{n} p_i$ and $\mathrm{Var}(S) = \sum_{i=1}^{n} p_i(1 - p_i)$.

In the special case that $p_i \equiv p$, $i = 1, \ldots, n$, the Poisson-Binomial distribution reduces to the Binomial distribution, $\mathrm{Bin}(n, p)$. Hence, expectation and variance of S also reduce to the well-known expressions np and $np(1 - p)$, respectively.

The R package poibin (Hong 2013), implements both exact and approximate methods for computing the cdf of the Poisson-Binomial distribution. It also provides the pmf, quantile function and random number generation for the Poisson-Binomial distribution.

Note that this distribution is not a *compound* Poisson distribution, and hence it is *not* to be interpreted as the distribution of a Poisson sum of Binomial distributions. Daskalakis, Diakonikolas, and Servedio (2012) remark that "It is believed that Poisson was the first to consider this extension of the Binomial distribution, and the distribution is sometimes referred to as 'Poisson's binomial distribution' " (An example of an actual compound distribution that we have seen earlier in this manuscript is the negative Binomial distribution, which can be written as a Poisson sum of lognormal distributions).

Mid-quantiles

Let X be a discrete random variable with distinct values v_j, $j = 0, \ldots, d$ (one usually will have $v_j \equiv j$). Let $P(X = v_j) = p_j$ and $\pi_j = \sum_{i=0}^{j-1} p_i + p_j/2$. The *mid-quantile* function for a probability γ is defined as (notation adapted from Ma *et al.* (2011))

$$Q_{\mathrm{mid}}(\gamma) = F^{-1}_{\mathrm{mid}}(\gamma) = \begin{cases} v_0 & \text{if } \gamma < p_0/2 \\ v_j & \text{if } \gamma = \pi_j,\ j = 0, \ldots, d \\ \lambda v_j + (1 - \lambda) v_{j+1} & \text{if } \gamma = \lambda \pi_j + (1 - \lambda)\pi_{j+1} \\ & 0 < \lambda < 1,\ j = 0, \ldots, d - 1 \\ v_d & \text{if } \gamma > \pi_d. \end{cases}$$

As a brief tutorial on he construction of mid-p-values, consider the random variable $X \sim \mathrm{Bin}(7, 0.35)$. The probability mass function of X is then given by the second row in Table 8, and the resulting values of π_j are given in the third row. This leads to the graphical representation of the mid-quantile function in Figure 8.

For the developments in this manuscript, the role of the random vector X will be taken by the random vector $N(k)$. Notably, we are not concerned with the distribution of $N(k)$ *over different* values of k, but with the possible, discrete, values of $N(k)$ for *fixed* k, with values in $v_0 = 0, \ldots, v_d = n$.

Table 8: Elements of the computation of mid-quantiles, for the binomial toy example

v_j=j	0	1	2	3	4	5	6	7
p_j	0.049	0.185	0.298	0.268	0.144	0.047	0.008	0.001
π_j	0.0245	0.1415	0.383	0.666	0.872	0.9675	0.9695	0.9995

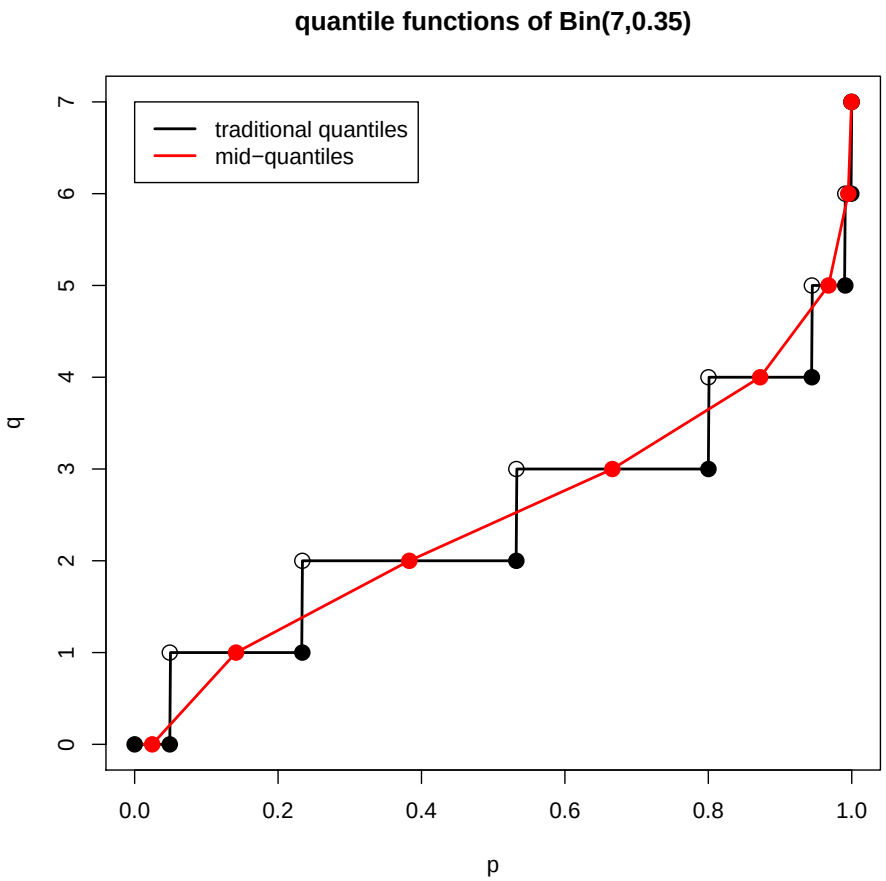

Figure 8: Illustration of (traditional) quantile and mid-quantile function for B(7,0.35) distribution.

R code for the computation of mid-quantiles from any probability mass function will be provided by the authors as supplementary material.

Density of Poisson model modified for digit preference

This probability mass function was used for generating data under digit preference in Section 2.2.

$$g(x,\mu,\phi) = \begin{cases} \frac{e^{-\mu}\mu^x}{x!} & x = 0,1,2 \\ \phi\frac{e^{-\mu}\mu^x}{x!} & x = 3,4,6,7,8,9,11\ldots \\ \frac{e^{-\mu}\mu^x}{x!} + (1-\phi)\left(\sum_{t=x-2}^{x-1}\frac{e^{-\mu}\mu^t}{t!} + \sum_{t=x+1}^{x+2}\frac{e^{-\mu}\mu^t}{t!}\right) & x = 5,10,15,\ldots \\ 0 & \text{otherwise} \end{cases}$$

Mid-quantile tables

Table 9: Mid-quantile table: Poisson model for digit preference data (Table 1)

k	$q_{0.025}(k)$	$N(k)$	$q_{0.975}(k)$	$m(k)$
0	0.00	1	2.49	0.39
1	0.05	5	7.21	3.01
2	5.05	9	18.05	10.95
3	16.58	20	36.13	25.80
4	33.20	32	58.49	45.32
5	49.40	106	78.72	63.59
6	59.12	69	90.39	74.30
7	59.19	64	90.48	74.38
8	50.81	57	80.42	65.14
9	37.93	35	64.49	50.69
10	24.69	55	47.34	35.47
11	13.97	22	32.30	22.54
12	6.56	9	20.80	13.10
13	2.27	4	12.87	6.98
14	0.18	4	7.84	3.43
15	0.00	6	4.81	1.54
16	0.00	2	2.97	0.62

Table 10: Mid-quantile table: Modified Poisson model for digit preference data (Table 1)

k	$q_{0.025}$	$N(k)$	$q_{0.975}$	$m(k)$
0	0.00	1	1.96	0.39
1	0.34	5	6.51	3.01
2	5.9	9	16.86	10.95
3	13.61	20	28.42	20.61
4	27.03	32	6.20	36.22
5	92.78	106	123.08	107.68
6	47.80	69	71.70	59.41
7	47.87	64	71.77	59.47
8	41.15	57	63.72	52.08
9	30.80	35	50.97	40.52
10	53.74	55	78.71	65.89
11	11.46	22	25.36	18.00
12	5.47	9	16.24	10.45
13	2.02	4	9.95	5.56
14	0.20	4	5.97	2.71
15	0.92	6	7.70	3.86
16	0.00	2	2.34	0.50

Table 11: Mid-quantile table: Poisson model for homicide data (Table 3)

k	$q_{0.025}$	$N(k)$	$q_{0.975}$	$m(k)$
0	670.35	1189	741.01	705.74
1	402.01	76	468.83	435.20
2	113.25	26	156.35	134.29
3	17.93	11	38.43	27.56
4	0.50	3	8.89	4.14
5	0.00	2	2.68	0.46
6	0.00	1	1.09	0.05

Table 12: Mid-quantile table: Geometric model for homicide data (Table 3)

k	$q_{0.025}$	$N(k)$	$q_{0.975}$	$m(k)$
0	415.35	1189	481.21	448.08
1	253.09	76	311.41	281.88
2	159.97	26	209.26	184.15
3	103.08	11	144.46	123.25
4	66.93	3	101.61	83.71
5	43.38	2	72.46	57.35
6	27.91	1	52.23	39.47

Table 13: Mid-quantile table: PIG model for homicide data (Table 3)

k	$q_{0.025}$	$N(k)$	$q_{0.975}$	$m(k)$
0	1165.33	1189	1205.59	1185.93
1	71.32	76	106.65	88.45
2	11.28	26	28.46	19.27
3	2.17	11	12.66	6.80
4	0.04	3	7.12	2.97
5	0.00	2	4.71	1.47
6	0.00	1	3.42	0.76

Table 14: Mid-quantile table: Negative Binomial (type II) model for homicide data (Table 3)

k	$q_{0.025}$	$N(k)$	$q_{0.975}$	$m(k)$
0	1167.58	1189	1207.62	1188.08
1	65.35	76	99.57	81.91
2	13.79	26	32.24	22.42
3	3.06	11	14.37	8.08
4	0.16	3	7.77	3.38
5	0.00	2	4.82	1.55
6	0.00	1	3.34	0.74

Table 15: Mid-quantile table: Negative Binomial (type I) model for homicide data (Table 3)

k	$q_{0.025}$	$N(k)$	$q_{0.975}$	$m(k)$
0	1167.57	1189	1207.62	1188.08
1	65.34	76	99.56	81.91
2	13.79	26	32.24	22.42
3	3.06	11	14.37	8.08
4	0.16	3	7.77	3.38
5	0.00	2	4.82	1.55
6	0.00	1	3.34	0.74

Table 16: Mid-quantile table: ZIP model for homicide data (Table 3).

k	$q_{0.025}$	$N(k)$	$q_{0.975}$	$m(k)$
0	1168.67	1189	1208.62	1189.13
1	56.58	76	88.97	72.22
2	20.32	26	41.64	19.27
3	4.94	11	17.81	10.77
4	0.21	3	7.90	3.50
5	0.00	2	3.81	0.97
6	0.00	1	1.96	0.27

Raw quantile band plots

These plots are included here to illustrate the advisability of subtraction of the raw information $N(k)$ by the median of their distribution. The supplementary material includes R code to produce such Raw Quantile Band plots.

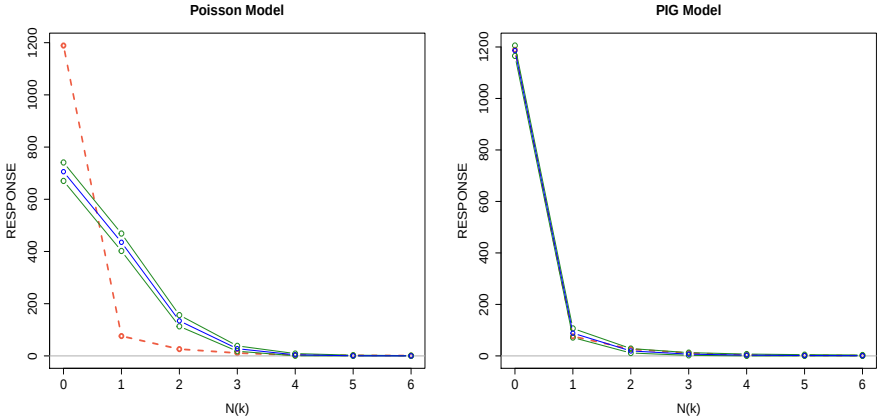

Figure 9: Raw Quantile Band plots for Poisson and PIG models for the data of Table 3 (Homicide data).

References

Agresti A (2002). *Categorical Data Analysis*. Wiley. ISBN: 978-0-471-45876-0.

Camarda C, Eilers P, Gampe J (2002). "Modelling General Patterns of Digit Preference." *Statistical Modelling*, **8**, 385–401.

Chen SX, Liu JS (1997). "Statistical Applications of the Poisson-binomial and Conditional Bernoulli Distributions." *Statistica Sinica*, **7**, 875–892.

da Silva W, Ribeiro AMT, Conceiçáo KS, Andrade MG, Louzada F (2018). "On Zero-modified Poisson-Sujatha Distribution to Model Overdispersed Count Data." *Austrian Journal of Statistics*, **47**, 1–19.

Daskalakis C, Diakonikolas I, Servedio RA (2012). "Learning Poisson Binomial Distributions." *Proceedings of the 44th Symposium on Theory of Computing*, pp. 709–728.

Eilers PH (2013). "Discussion: The Beauty of Expectiles." *Statistical Modelling*, **13**(2), 317–322. Doi: 10.1177/1471082X13494313.

Franck W (1986). "*P*–values for Discrete Test Statistics." *Biometrical Journal*, **28**, 403–406.

Hong Y (2013). *poibin: The Poisson Binomial Distribution.* R package version 1.2., URL https://CAN.-project.org/package=poibin.

Irwin JO (1959). "On the Estimation of the Mean of a Poisson Distribution from a Sample with the Zero Class Missing." *Biometrics*, **15**, 324–326.

Ma Y, Genton M, Parzen E (2011). "Quantiles of Discrete Distributions." *Annals of the Institute of Statistical Mathematics*, **63**, 227–243.

McCullagh P, Nelder JA (1989). *Generalized Linear Models.* Chapman and Hall. ISBN 978-1482203530.

Myers R (2002). "Errors and Bias in the Reporting of Ages in Census Data." *Transactions of the Actuarial Society of America*, **41**(2 (104)), 395–415.

Oliveira M, Einbeck J, Higueras M, Ainsbury E, Puig P, Rothkamm K (2016). "Zero-inflated Regression Models for Radiation–induced Chromosome Aberration Data: A Comparative Study." *Biometrical Journal*, **58**, 259–279.

Plackett RL (1953). "The Truncated Poisson Distribution." *Biometrics*, **9**, 485–488.

Puig P, Barquineiro JF (2011). "An Application of Compound Poisson Modelling to Biological Dosimetry." *Proceedings of the Royal Society A*, **467**, 897–910.

Ridout MS, CB D (1992). "Generalized Linear Models for Positive Count Data." *Revista de Matemática e Estatística*, **10**, 139–148.

Rigby A, Stasinopoulos DM (2005). "Generalized Additive Models for Location, Scale and Shape (with Discussion)." *Appl. Statist. (JRSSC)*, **54**, 507–554.

Wilson P, Einbeck J (2018). "A New and Intuitive Test for Zero Modification." *Statistical Modelling*, **19**(4), 341–361. Doi: 1471082X1876227.

Yee TW (2010). "The VGAM Package for Categorical Data Analysis." *Journal of Statistical Software*. R package version 1.2., URL http://www.jstatsoft.org/v32/i10/.

Affiliation:

Paul Wilson
School of Mathematics and Computer Science
Faculty of Science and Engineering
University of Wolverhampton
Wulfruna Street
Wolverhampton
WV1 1LY
United Kingdom
Telephone: +44/1902/321444
pauljwilson@wlv.ac.uk

Jochen Einbeck
Department of Mathematical Sciences
Durham University
South Road
Science Laboratories
Durham City
DH1 3LE
United Kingdom
Telephone: +44/191/3343125
jochen.einbeck@durham.ac.uk

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 50 *Submitted:* 2019-01-03
January 2021 *Accepted:* 2019-04-01